Econometrics: What's It All About, Alfie?

Using sample data on observable variables to learn about economic relationships, the functional relationships among economic variables.

Econometrics consists mainly of:

- ***estimating*** economic relationships from sample data

- ***testing hypotheses*** about how economic variables are related

    - the ***existence*** **of relationships** between variables
    - the ***direction*** **of the relationships** between one economic variable -- the dependent or outcome variable -- and its hypothesized observable determinants
    - the ***magnitude*** **of the relationships** between a dependent variable and the independent variables that are thought to determine it.

Sample data consist of ***observations*** **on** ***randomly selected*** **members of populations** of economic agents (individual persons, households or families, firms) or other units of observation (industries, provinces or states, countries).

Example 1

We wish to investigate empirically the determinants of households' food expenditures, in particular the relationship between households' food expenditures and households' incomes.

Sample data consist of **a random sample of 38 households** from the population of all households. For each household in the random sample, we have observations on three observable variables:

foodexp = annual food expenditure of household, thousands of dollars per year

income = annual income of household, thousands of dollars per year

hhsize = household size, number of persons in household

```
. list foodexp income hhsize

       foodexp     income     hhsize
  1.    15.998     62.476          1
  2.    16.652     82.304          5
  3.    21.741     74.679          3
  4.     7.431     39.151          3
  5.    10.481     64.724          5
  6.    13.548     36.786          3
  7.    23.256     83.052          4
  8.    17.976     86.935          1
  9.    14.161     88.233          2
 10.     8.825     38.695          2
 11.    14.184     73.831          7
 12.    19.604     77.122          3
 13.    13.728     45.519          2
 14.    21.141     82.251          2
 15.    17.446     59.862          3
 16.     9.629     26.563          3
 17.    14.005     61.818          2
 18.      9.16     29.682          1
 19.    18.831     50.825          5
 20.     7.641     71.062          4
 21.    13.882      41.99          4
 22.      9.67     37.324          3
 23.    21.604     86.352          5
 24.    10.866     45.506          2
 25.     28.98     69.929          6
 26.    10.882     61.041          2
 27.    18.561     82.469          1
 28.    11.629     44.208          2
 29.    18.067     49.467          5
 30.    14.539     25.905          5
 31.    19.192     79.178          5
 32.    25.918     75.811          3
 33.    28.833     82.718          6
 34.    15.869     48.311          4
 35.     14.91     42.494          5
 36.      9.55     40.573          4
 37.    23.066     44.872          6
 38.    14.751     27.167          7

. describe

Contains data from foodexp.dta
  obs:            38
 vars:             3                         7 Sep 2000 23:30
 size:           608 (99.9% of memory free)
-------------------------------------------------------------------------------
   1. foodexp   float  %9.0g                 food expenditure, thousands $
                                               per yr
   2. income    float  %9.0g                 household income, thousands $
                                               per yr
   3. hhsize    float  %9.0g                 household size, persons per hh
-------------------------------------------------------------------------------
Sorted by:
     Note:  dataset has changed since last saved
```

```
. summarize

Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+-----------------------------------------------------
 foodexp |      38    15.95282   5.624341      7.431      28.98
  income |      38    58.44434   19.93156     25.905     88.233
  hhsize |      38    3.578947   1.702646          1          7
```

<u>Question</u>: What relationship generated these sample data? What is the data generating process?

<u>Answer</u>: We postulate that **each population value of foodexp**, denoted as **$foodexp_i$**, is generated by a relationship of the form:

$$foodexp_i = f(income_i, hhsize_i) + u_i \qquad \Leftarrow \text{ the } \textbf{\textit{population}} \textbf{ regression equation}$$

where

$food\,exp_i$ = the ***dependent*** **or** ***outcome*** **variable** we are trying to explain
= the annual food expenditure of household i (thousands of $ per year)

$income_i$ = on ***independent*** **or** ***explanatory*** **variable** that we think might explain the dependent variable $food\,exp_i$
= the annual income of household i (thousands of $ per year)

$hhsize_i$ = a second ***independent*** **or** ***explanatory*** **variable** that we think might explain the dependent variable $food\,exp_i$
= household size, measured by the number of persons in the household

$f(income_i, hhsize_i)$ = a **population regression function** representing the systematic relationship of $food\,exp_i$ to the independent or explanatory variables $income_i$ and $hhsize_i$;

u_i = an ***unobservable random*** **error term** representing **all *unknown* and *unmeasured* variables** that determine the individual population values of $food\,exp_i$

<u>Question</u>: What mathematical form does the population regression function $f(\text{income}_i, \text{hhsize}_i)$ take?

<u>Answer</u>: We hypothesize that the **population regression function** -- or **PRF --** is a linear function:

$$f(\text{income}_i, \text{hhsize}_i) = \beta_0 + \beta_1 \text{income}_i + \beta_2 \text{hhsize}_i$$

<u>Implication</u>: The **population regression equation** -- the **PRE --** is therefore

$$\begin{aligned} \text{food}\exp_i &= f(\text{income}_i, \text{hhsize}_i) + u_i \\ &= \beta_0 + \beta_1 \text{income}_i + \beta_2 \text{hhsize}_i + u_i \end{aligned}$$

- **Observable Variables:**

  $\text{foodexp}_i \equiv$ the value of the dependent variable foodexp for the i-th household
  $\text{income}_i \equiv$ the value of the independent variable income for the i-th household
  $\text{hhsize}_i \equiv$ the value of the independent variable hhsize for the i-th household

- **Unobservable Variable:**

  $u_i \equiv$ the value of the random error term for the i-th household in the population

- **Unknown Parameters:** the **regression coefficients β_0, β_1 and β_2**

  β_0 = the ***intercept* coefficient**
  β_1 = the ***slope* coefficient on income_i**
  β_2 = the ***slope* coefficient on hhsize_i**

  The **population values** of the regression coefficients **β_0, β_1 and β_2** are ***unknown***.

Example 2

We wish to investigate empirically the determinants of paid workers' wage rates. In particular, we want to investigate whether male and female workers with the same characteristics on average earn the same wage rate.

Sample data consist of **a random sample of 526 paid workers** from the 1976 US population of all paid workers in the employed labour force.

For each paid worker in the random sample, we have observations on six observable variables:

wage	=	average hourly earnings of paid worker, dollars per hour
ed	=	years of education completed by paid worker, years
exp	=	years of potential work experience of paid worker, years
ten	=	tenure, or years with current employer, of paid worker, years
female	=	1 if paid worker is female, = 0 otherwise
married	=	1 if paid worker is married, = 0 otherwise

```
. describe

Contains data from wage1.dta
  obs:           526
 vars:             6                          16 Apr 2000 16:18
 size:        94,680 (90.7% of memory free)
-------------------------------------------------------------------------------
   1. wage      float  %9.0g                  average hourly earnings, $/hour
   2. ed        float  %9.0g                  years of education
   3. exp       float  %9.0g                  years of potential work
                                                experience
   4. ten       float  %9.0g                  tenure = years with current
                                                employer
   5. female    float  %9.0g                  =1 if female, =0 otherwise
   6. married   float  %9.0g                  =1 if married, =0 otherwise
```

```
. list wage ed exp ten female married

          wage         ed         exp         ten      female    married
  1.     21.86         12          24          16           0          1
  2.       5.5         12          18           3           0          0
  3.      3.75          2          39          13           0          1
  4.        10         12          31           2           0          1
  5.       3.5         13           1           0           0          0
  6.      6.67         12          35          10           0          0
  7.      3.88         12          12           3           0          1
  8.      5.91         12          14           6           0          1
  9.       5.9         12          14           7           0          1
 10.        10         17           5           3           0          1
 11.      4.55         16          34           2           0          1
 12.        10          8           9           0           0          1
 13.         6         13           8           0           0          1
 14.         5          9          31           9           0          1
 15.       4.5         12          13           0           0          1
 16.      5.43         14          10           3           0          1
 17.      2.83         10           1           0           0          0
 18.       6.8         12          14          10           0          1
 19.      6.76         12          19           3           0          1
 20.      4.51         12           5           2           0          0

(output omitted)

507.      6.15         12          35          12           1          0
508.      11.1         15           1           4           1          0
509.      3.35         15           3           1           1          1
510.         5         12           7           3           1          0
511.      3.35          7          35           0           1          0
512.      6.25         12          13           0           1          1
513.      3.06         12          14          10           1          0
514.       5.9         12           9           7           1          1
515.       8.1         12          38           3           1          1
516.     14.58         18          13           7           1          0
517.      9.42         14          23           0           1          1
518.      9.68         13          16          16           1          0
519.       8.6         16           3           2           1          0
520.         3         12          38           0           1          1
521.      3.33         12          45           4           1          1
522.         4         12          22          11           1          1
523.      2.75         13           1           2           1          0
524.         3         16          19          10           1          1
525.       2.9          8          49           6           1          0
526.      3.18         12           5           0           1          1

. summarize wage ed exp ten female married

Variable |      Obs        Mean   Std. Dev.       Min        Max
---------+-----------------------------------------------------
    wage |      526    5.896103   3.693086        .53      24.98
      ed |      526    12.56274   2.769022          0         18
     exp |      526    17.01711   13.57216          1         51
     ten |      526    5.104563   7.224462          0         44
  female |      526    .4790875    .500038          0          1
 married |      526     .608365   .4885804          0          1
```

Question: **What relationship generated these sample data?** What is the data generating process?

Answer: We postulate that **each population value of wage**, denoted as **$\mathbf{wage_i}$**, is generated by a ***population*** **regression equation** of the form:

$$wage_i = f(ed_i, exp_i, ten_i, female_i, married_i) + u_i$$

where:

$wage_i$ = the ***dependent*** **or** ***outcome*** **variable** we are trying to explain
= the average hourly earnings of paid worker i (dollars per hour)

ed_i = one ***independent*** **or** ***explanatory*** **variable** that we think might explain the dependent variable $wage_i$
= the years of education completed by paid worker i (years)

exp_i = a second ***independent*** **or** ***explanatory*** **variable** that might explain $wage_i$
= the potential work experience accumulated by paid worker i (years)

ten_i = a third ***independent*** **or** ***explanatory*** **variable** that might explain $wage_i$
= tenure, years with current employer, of paid worker i (years)

$female_i$ = a fourth ***independent*** **or** ***explanatory*** **variable** that might affect $wage_i$
= 1 if paid worker i is female, = 0 otherwise

$married_i$ = a fifth ***independent*** **or** ***explanatory*** **variable** that we think might explain the dependent variable $wage_i$
= 1 if paid worker i is married, = 0 otherwise

$f(ed_i, exp_i, ten_i, female_i, married_i)$
= a **population regression function** representing the systematic relationship of $wage_i$ to the independent variables ed_i, exp_i, ten_i, $female_i$ and $married_i$

u_i = an ***unobservable random*** **error term** representing **all** ***unknown*** **variables and** ***unmeasured*** that determine the individual population values of $wage_i$

<u>Question</u>: What mathematical form does the population regression function, or PRF, $f(ed_i, \cdots, married_i)$ take?

<u>Answer</u>: We hypothesize that the **population regression function** -- or **PRF** -- is a ***linear* function**.

$$f(ed_i, \cdots, married_i) = \beta_0 + \beta_1 ed_i + \beta_2 \exp_i + \beta_3 ten_i + \beta_4 female_i + \beta_5 married_i$$

<u>Implication</u>: The **population regression equation** -- the **PRE** -- is therefore

$$\begin{aligned} wage_i &= f(ed_i, \exp_i, ten_i, female_i, married_i) + u_i \\ &= \beta_0 + \beta_1 ed_i + \beta_2 \exp_i + \beta_3 ten_i + \beta_4 female_i + \beta_5 married_i + u_i \end{aligned}$$

- **Observable Variables:**

  $wage_i \equiv$ the value of the dependent variable wage for the i-th employee
  $ed_i \equiv$ the value of the independent variable ed for the i-th employee
  $exp_i \equiv$ the value of the independent variable exp for the i-th employee
  $ten_i \equiv$ the value of the independent variable ten for the i-th employee
  $female_i \equiv$ the value of the independent variable female for the i-th employee
  $married_i \equiv$ the value of the independent variable married for the i-th employee

- **Unobservable Variable:**

  $u_i \equiv$ the value of the random error term for the i-th paid worker in the population

- **Unknown Parameters:** the **regression coefficients β_0, β_1, β_2, β_3, β_4 and β_5**

  β_0 = the ***intercept* coefficient**
  β_1 = the ***slope* coefficient on ed_i**
  β_2 = the ***slope* coefficient on exp_i**
  β_3 = the ***slope* coefficient on ten_i**
  β_4 = the ***slope* coefficient on $female_i$**
  β_5 = the ***slope* coefficient on $married_i$**

<u>Our task</u>: To learn how to compute from sample data reliable estimates of the regression coefficients β_0, β_1, β_2, β_3, β_4 and β_5.

The Four Elements of Econometrics

Data

Collecting and coding the sample data, the raw material of econometrics.

Most economic data is ***observational*, or *non-experimental*, data** (as distinct from ***experimental*** **data** generated under controlled experimental conditions).

Specification

Specification of the *econometric model* that we think (hope) generated the sample data -- that is, specification of the **data generating process** (or DGP).

An ***econometric model*** consists of **two components**:

1. An ***economic model***: specifies the ***dependent* or *outcome* variable** to be explained and the ***independent* or *explanatory* variables** that we think are related to the dependent variable of interest.
   - Often suggested or derived from economic theory.
   - Sometimes obtained from informal intuition and observation.

2. A ***statistical model***: specifies the statistical elements of the relationship under investigation, in particular the ***statistical properties* of the *random* variables** in the relationship.

Estimation

Consists of **using the assembled *sample data*** on the ***observable* variables** in the model **to compute *estimates*** of the **numerical values** of all the **unknown parameters** in the model.

Inference

Consists of **using the parameter estimates** computed from sample data **to test hypotheses** about the **numerical values** of the **unknown *population* parameters** that describe the behaviour of the population from which the sample was selected.

Scientific Method

The collection of principles and processes necessary for scientific investigation, including:

1. rules for concept formation

2. rules for conducting observations and experiments

3. rules for validating hypotheses by observations or experiments

Econometrics is that branch of economics -- the dismal science -- which is concerned with items 2 and 3 in the above list.

Recap

We have considered **two examples** of what are generically called ***linear regression equations*** or ***linear regression models***.

Example 1 -- a linear regression model for household food expenditure:

$$\text{food exp}_i = \beta_0 + \beta_1 \text{income}_i + \beta_2 \text{hhsize}_i + u_i$$

Example 2 -- a linear regression model for paid workers' wage rates:

$$\text{wage}_i = \beta_0 + \beta_1 \text{ed}_i + \beta_2 \exp_i + \beta_3 \text{ten}_i + \beta_4 \text{female}_i + \beta_5 \text{married}_i + u_i$$

Regression analysis has two fundamental tasks:

1. **Estimation: computing from *sample* data** reliable ***estimates*** of the ***numerical* values of the regression coefficients β_j (j = 0, 1, …, K)**, and hence of the population regression function.

2. **Inference:** using ***sample estimates*** **of the regression coefficients β_j (j = 0, 1, …, K)** to **test hypotheses** about the ***population*** **values** of the **unknown regression coefficients** -- i.e., to ***infer*** **from *sample*** **estimates** the **true *population* values of the regression coefficients** *within specified margins of statistical error*.